

# Let's make cheap reference sequences

(work in progress)

David Jaffe

SFAF – May 28, 2015



**BROAD**  
INSTITUTE



**10X**<sup>TM</sup>  
*DE NOVO*

# 10X: the next generation *de novo*

## Goals

- Normal human genomes and tumors
- Everything

4 requirements  
for  
genome analysis

R1: don't use more DNA than you have

---



ng DNA?

R2: don't use more money than you have



CHEAP!



Too much  
for one ant

**Ants of New York City**

|  |  |   |   |
|--|--|---|---|
| <br>Carpenter Ant     | <br>Lasius Ant  | <br>Pavement Ant   |   |
| <br>Odorous House Ant | <br>Crazy Ant   | <br>Winter Ant     | <br>Field Ant        |
| <br>Asian Needle Ant  | <br>Winnow Ant  | <br>Big Headed Ant | <br>Little Black Ant |
| <br>Thief Ant         | <br>Acrobat Ant | <br>Honeyrump Ant  | <br>Ant Biology      |

**and then the planet**

R3: see all the bases in the DNA

---

| bases to see<br>↓     | <i>de novo</i><br>world | resequencing<br>world |
|-----------------------|-------------------------|-----------------------|
| DNA not in reference? | ✓                       | nope                  |
| heterozygous sites?   | nope                    | ✓                     |

Give us both!

## R4: resolve structure

---

### Reveal the sequence of individual chromosomes

- Phase
- Pull apart segmental duplications
- Reveal structural variation



## Data enabling our requirements

---

|                  |                 |                    |                            |
|------------------|-----------------|--------------------|----------------------------|
| R1<br>DNA<br>low | R2<br>\$<br>low | R3<br>all<br>bases | R4<br>resolve<br>structure |
|------------------|-----------------|--------------------|----------------------------|

It's got to be short reads

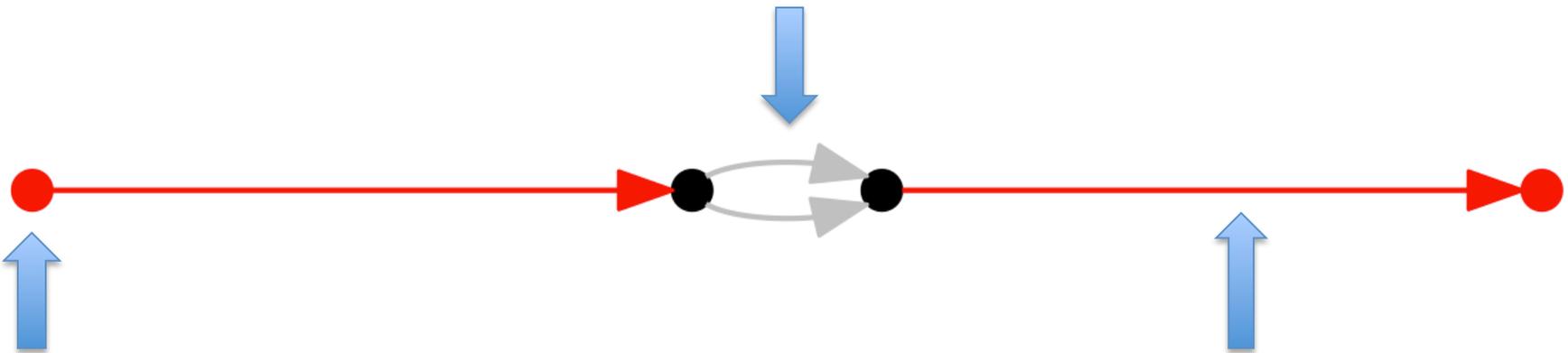
|   |  |
|---|--|
| <u>maximum local quality</u><br><br>500 ng DNA<br><br>one PCR-free library<br><br>2x250 reads | <u>maximum global quality</u><br><br>1 ng DNA<br><br>one 10X library<br><br>2x88 reads |
|---|--|

I'LL START HERE: DISCOVER *DE NOVO*

# DISCOVAR *de novo* graph captures variant positions

---

'Bubbles' represent alternative paths, *typically* heterozygous sites



Only showing *part* of the graph.  
Red vertices: the graph continues on....

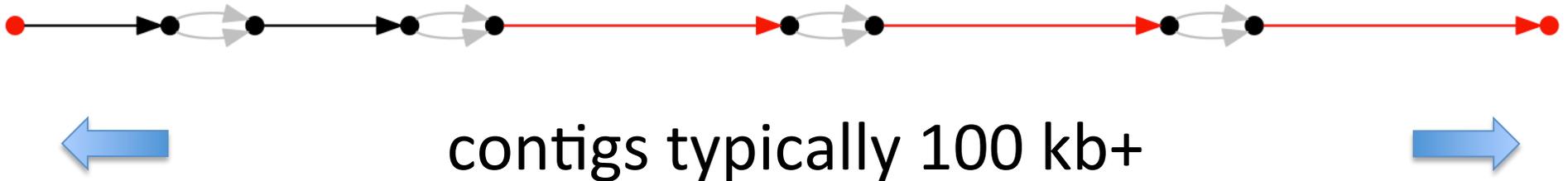
Each edge represents a DNA sequence

Edges color-coded by length: gray < black < red < magenta

## DISCOVAR *de novo* assemblies

---

Mostly very long lines with bubbles



Show chromosome counts

~6 tumor  
chromosomes      ~2 normal  
chromosomes



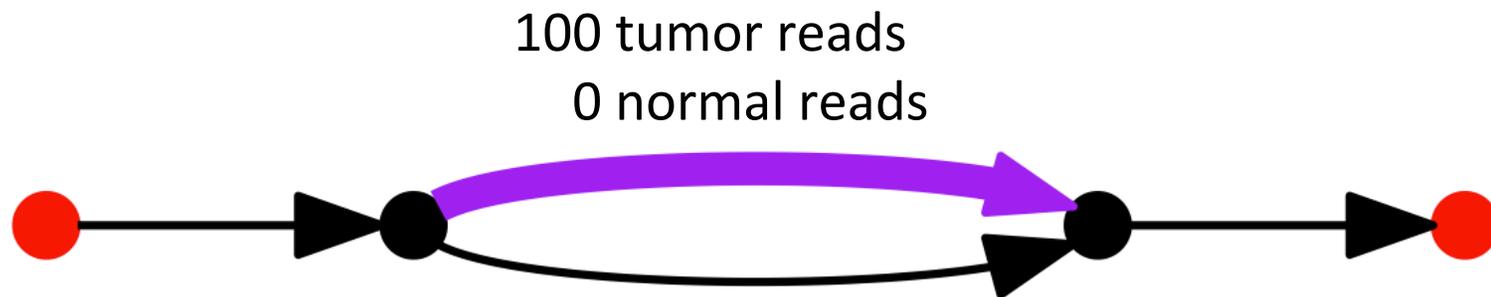
1 [6.00x;2.06x] (3.5 kb)

The diagram shows a horizontal red line representing a contig, starting with a red dot on the left and ending with a red dot on the right. The text "1 [6.00x;2.06x] (3.5 kb)" is centered above the line.

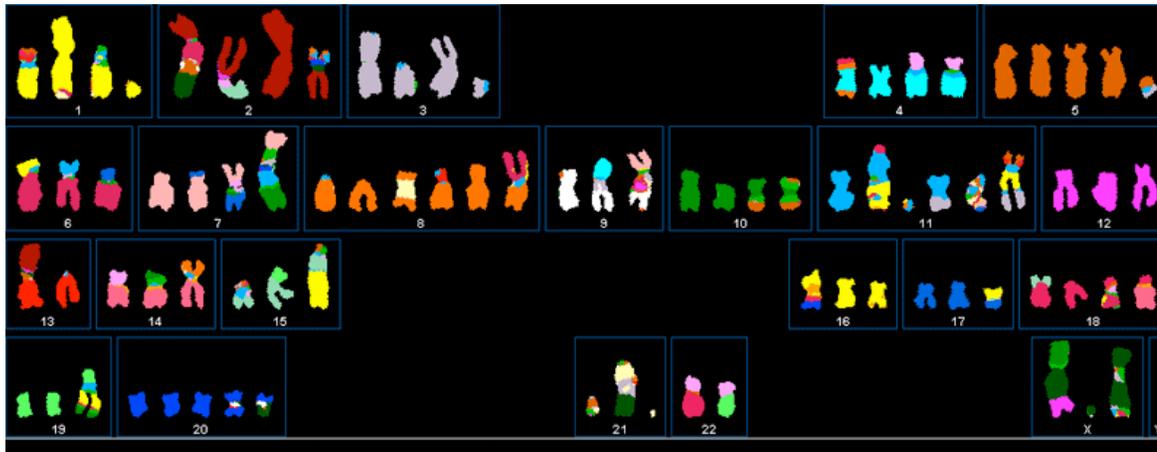
Find somatic mutations, of ALL types?

---

Look for tumor-only edges!

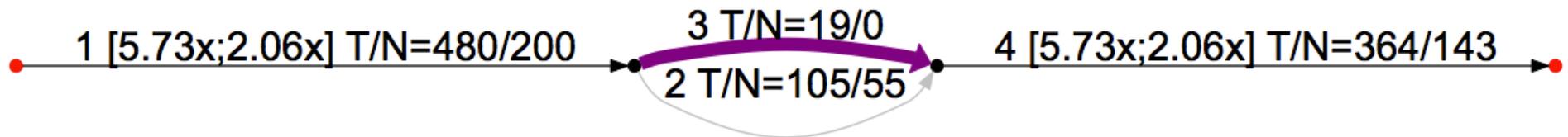


Apply to



HCC1143  
messed up  
breast cancer  
cell line  
→ N50 contig  
**115 kb**

? S=10:11M D=1



Not found by standard methods.

? ALIGN 2,3

```
GCAGTTGATAAATTGGGGTCAGAGGAGCTTTGTGTCTTCAGTGTGACATCAGCAGAATAGAAGAACGTTCCATCCATTTGC
GCAGTTGATAAATTGGGGTCAGAGGAGCTTTGTGTCTTCAGTGTGACATCAGCAGAATAGAAGAACGTTCCATCCATTTGC
```

(80 matching bases)

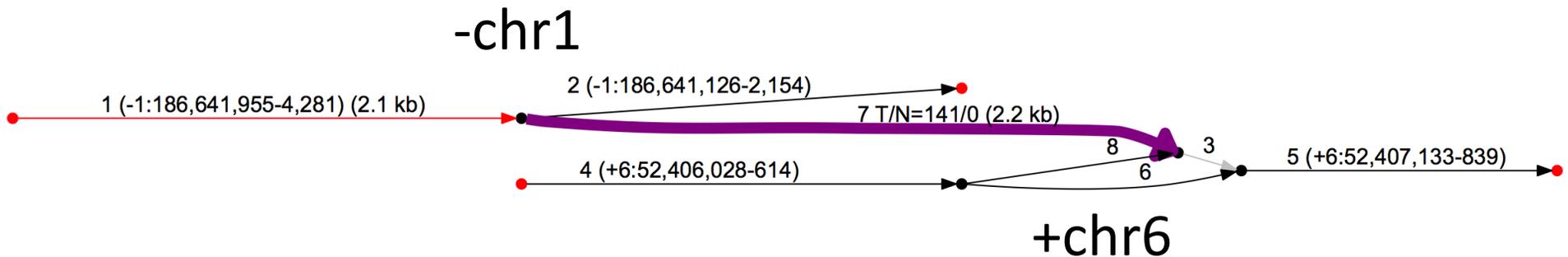
||

```
GATCATTGCTTCATATTTACTTAACAAATTTTCTATGTATTCTTTCCATCTTCCAGAGCAATATGGGATGTTTCTTAATT
GATCATTGCTTCATATTTACTTAACAAATTTTCTATGTA CTTTCCATCTTCCAGAGCAATATGGGATGTTTCTTAATT
```

(160 matching bases)

# DISCOVAR finds crazy-quilt rearrangements

---



Not found by standard methods.

Decomposition of purple edge into  
13 segments from all across the genome:

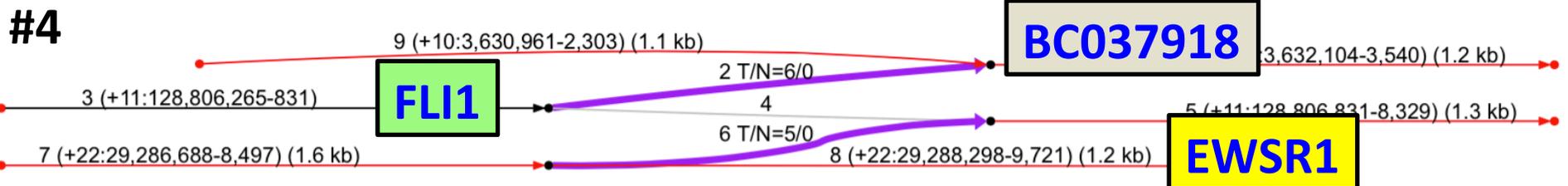
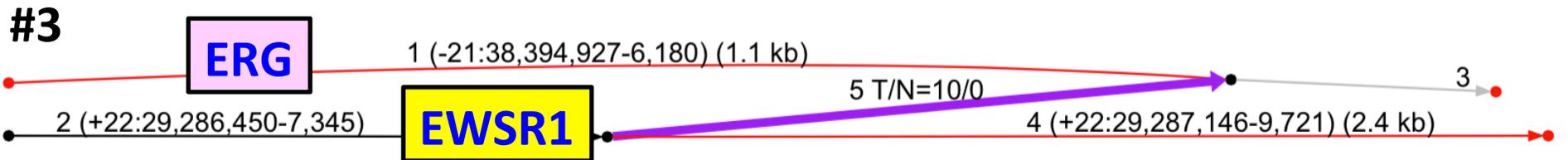
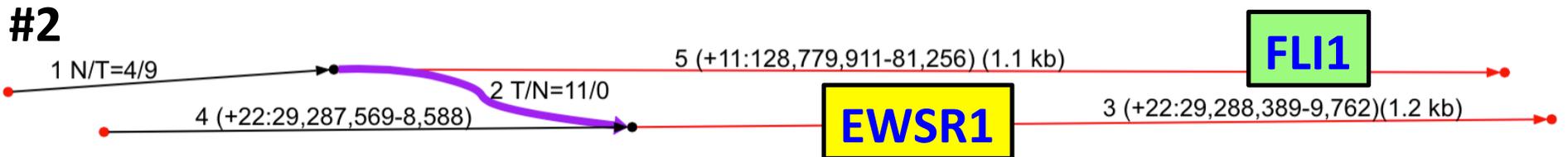
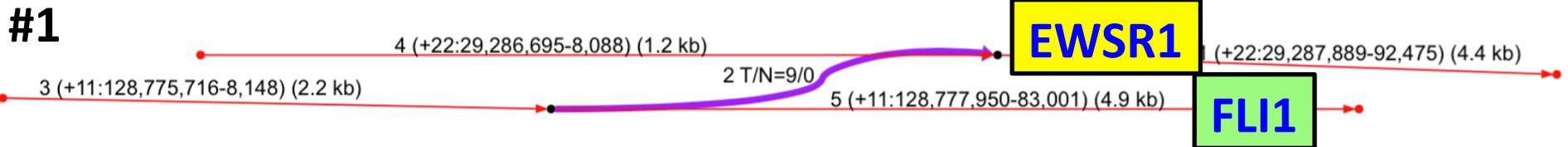


# A look at 17x Ewing sarcoma assemblies from patient samples

NATURE · VOL 359 · 10 SEPTEMBER 1992

**Gene fusion with an *ETS* DNA-binding domain caused by chromosome translocation in human tumours**

EWING'S sarcoma and related subtypes of primitive neuroectodermal tumours share a recurrent and specific t(11; 22) (q24; q12) chromosome translocation<sup>1-8</sup>, the breakpoints of which have recently been cloned<sup>9</sup>. Phylogenetically conserved restriction fragments in the vicinity of *EWSR1* and *EWSR2*, the genomic regions



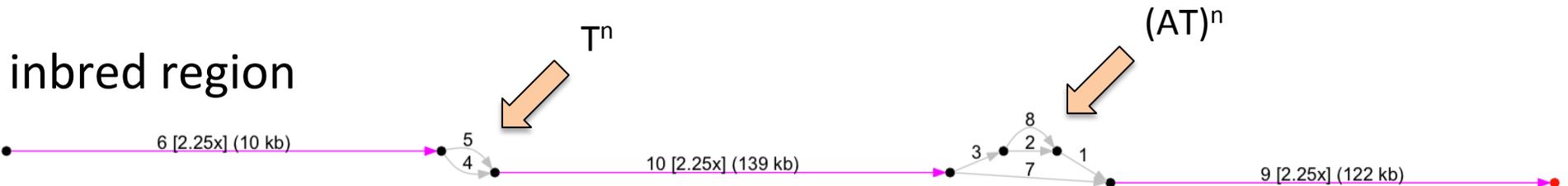


nonhumans matter too

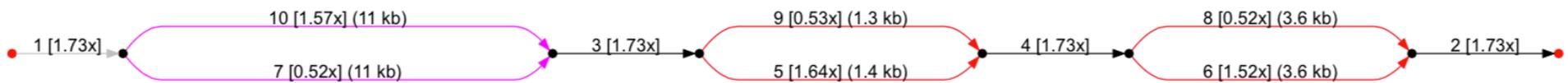


*Anopheles arabiensis*  
N50 contig size 21 kb

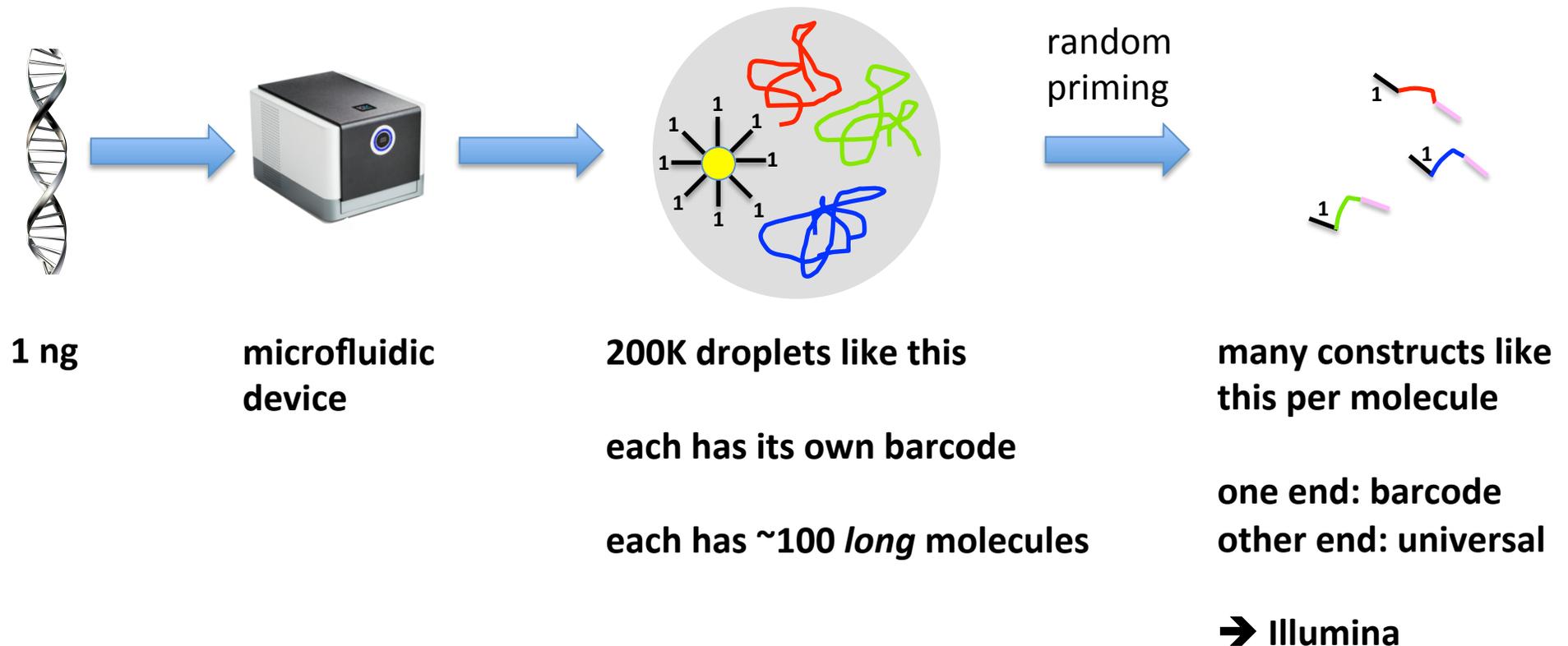
largely inbred progeny of cross



heterozygous region, one chromosome of four (??) different



# 10X Genomics → Parallel Single Haplotype Libraries



Strategy: use these data to resolve DISCOVAR *de novo* assembly graph

Remarkably, it works!

---

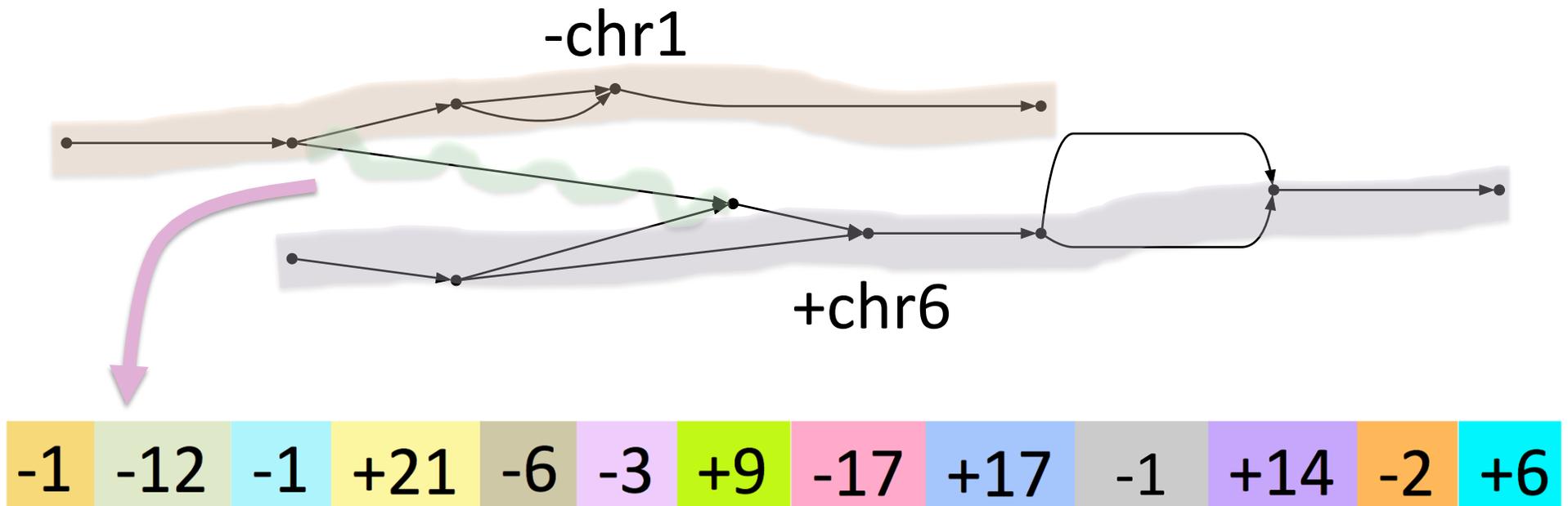


160 kb on *one* chromosome

10X power  $\approx$  giant perfect reads

## Crazy-quilt rearrangement – revisited

---



Supported by 134 tumor reads, 0 normal reads

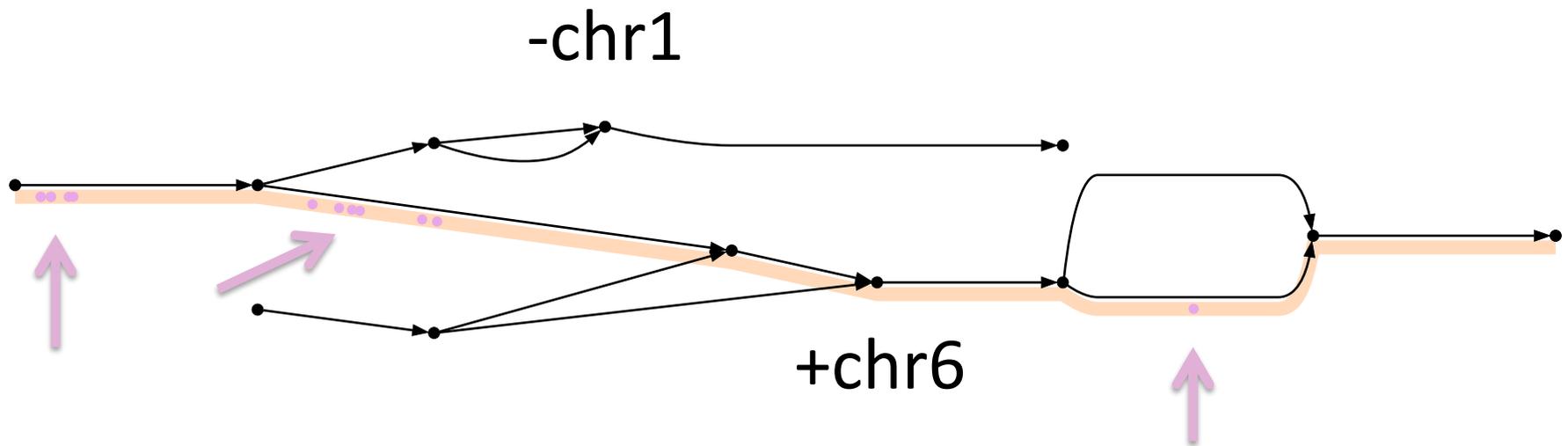
Not seen by standard methods. Should we believe it?

# Now add 10X data: uniquely placed reads define path

---

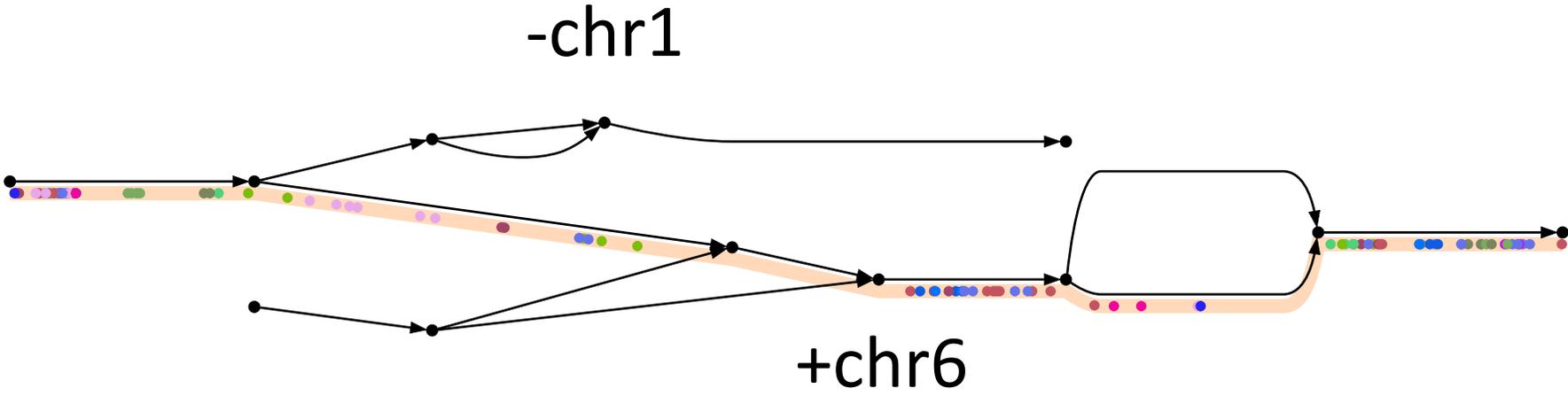
Reads from *one* barcode define a path

Only **unique** placements on entire assembly shown



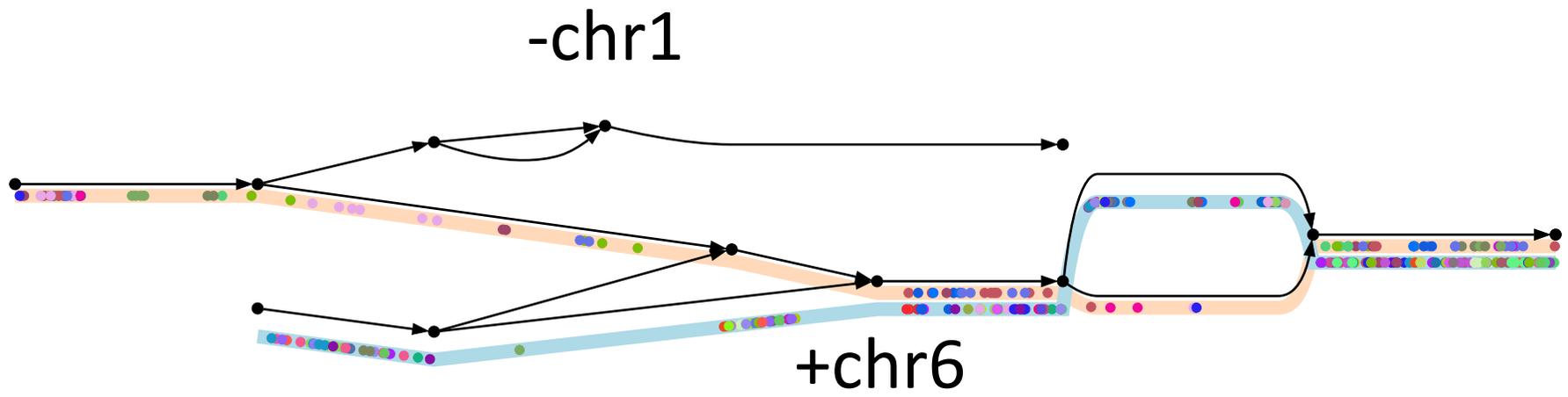
# Multiple barcodes vouch for the path

---



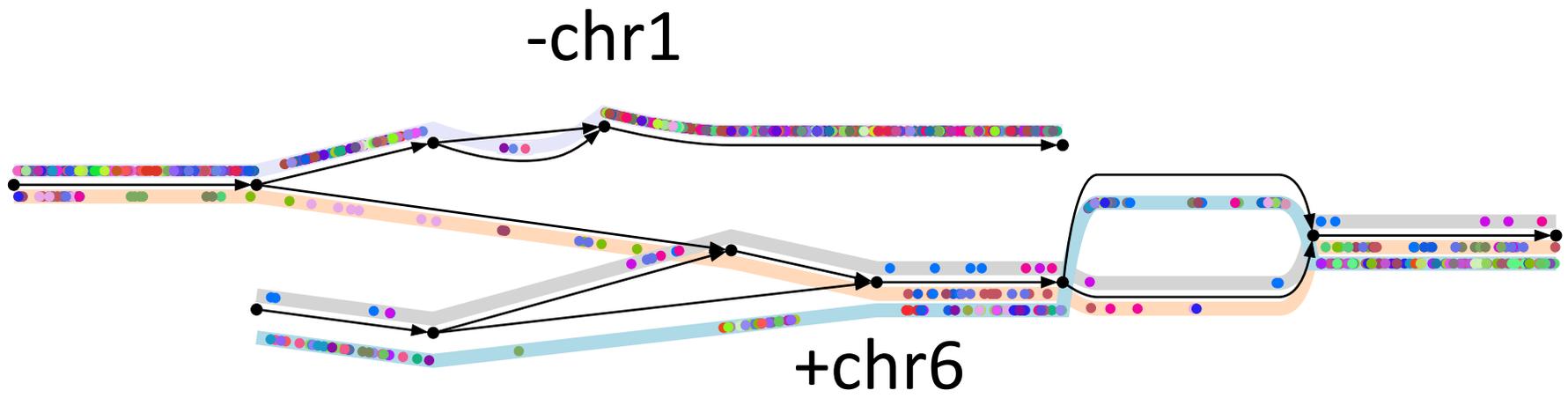
# Another chromosome is uniquely determined

---



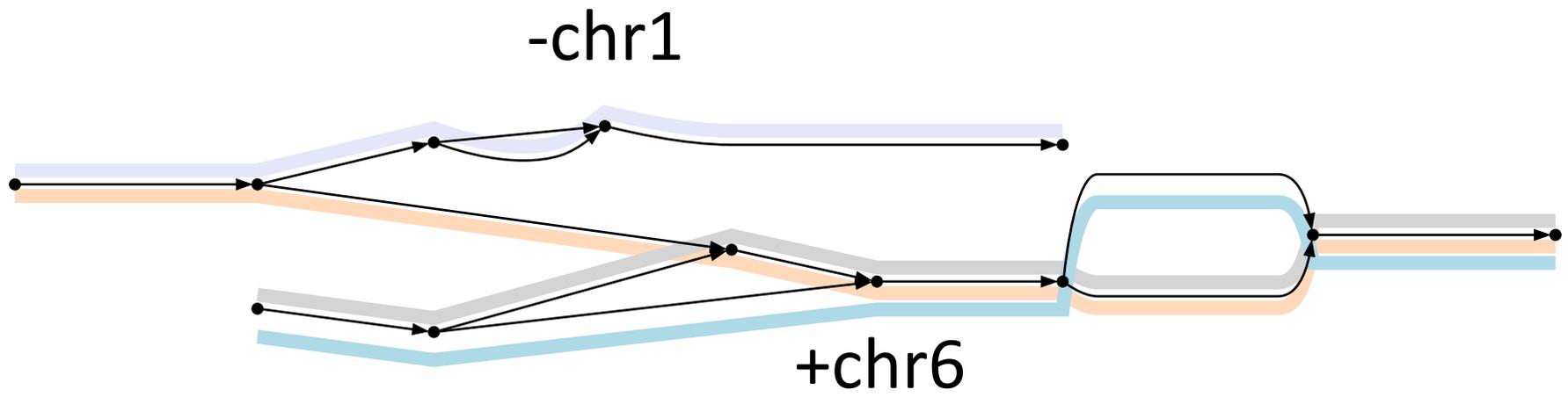
# All four tumor chromosomes completely determined

---



All four tumor chromosomes completely determined

---



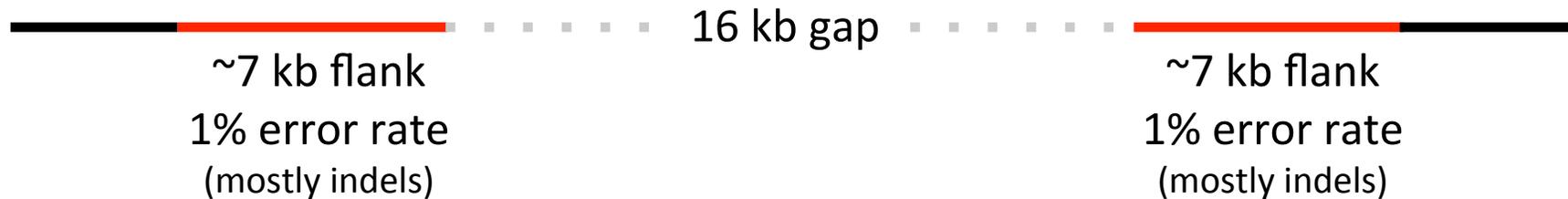
# 10X resolves segmental duplications

---

38 kb

*chr3:0.574M-0.612M*

## PacBio assembly of CHM1



## DISCOVAR *de novo* assembly

10 contigs larger than 1kb (blue)



Smaller connecting contigs. Intertwined with other chromosomes.

## DISCOVAR *de novo* + 10X assembly



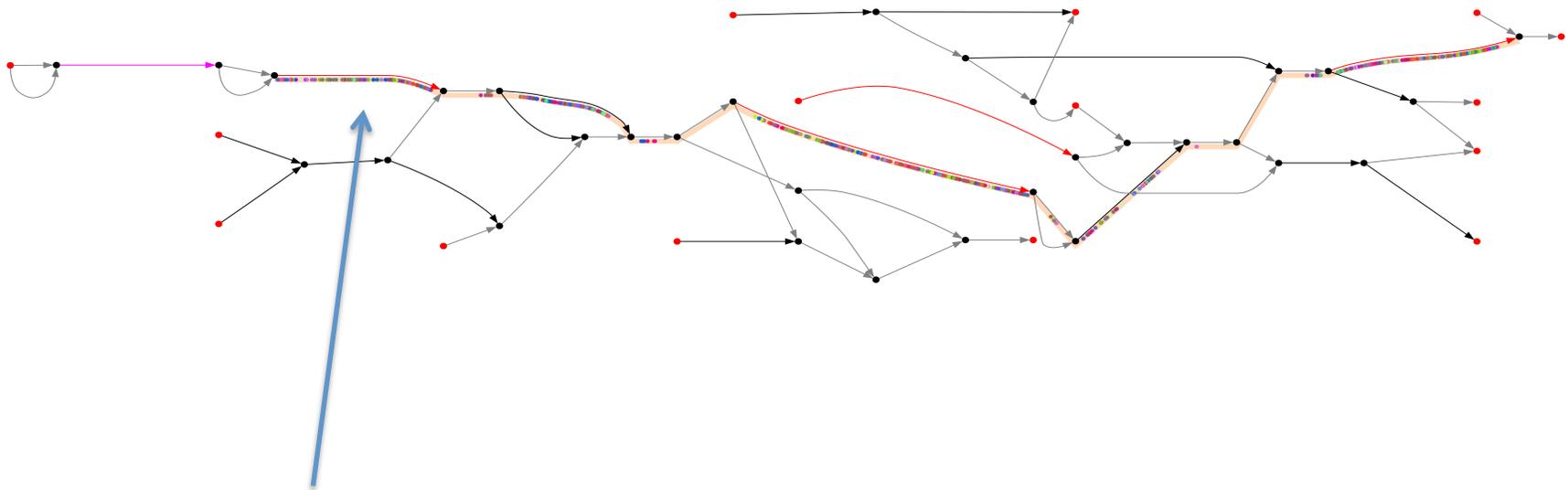
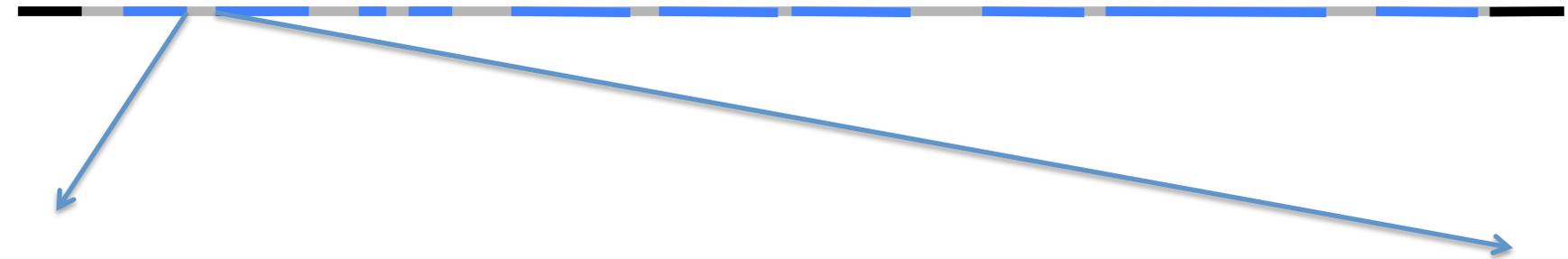
Completely resolved!

# How one leg of this duplication gets resolved

---

DISCOVAR assembly

10 contigs larger than 1kb (blue)



seed here and use 10X to walk forward

# 10X transformative for *de novo* assembly

---

The data satisfy our requirements

|                  |                 |                    |                            |
|------------------|-----------------|--------------------|----------------------------|
| R1<br>DNA<br>low | R2<br>\$<br>low | R3<br>all<br>bases | R4<br>resolve<br>structure |
|------------------|-----------------|--------------------|----------------------------|

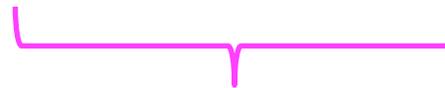


One strategy

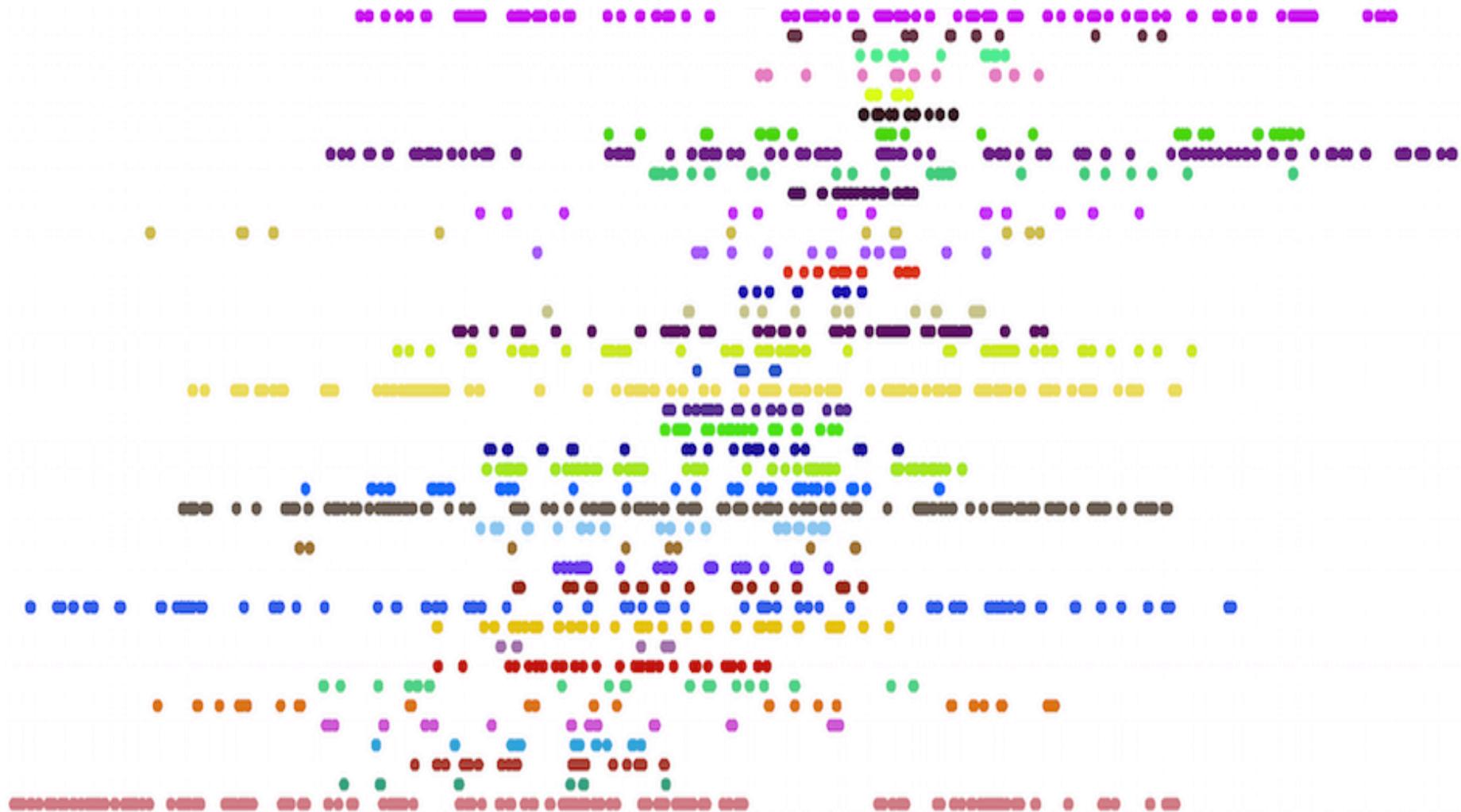
2x250 PCR-free  
DISCOVAR *de novo*



10X  
walk the graph



10X may be enough!



# Thank you

---



Neil Weisenfeld  
Ted Sharpe  
Iain MacCallum

## Broad Genomics Platform

Ryan Hegarty  
Tim DeSmet  
Laurie Holmes  
Katherine Sullivan

## Illumina

Feng Chen  
Eric Jaeger  
Geoff Smith  
Gary Schroth

## Broad Infectious Disease

Dan Neafsey

## Broad Technologies Lab

Georgia Giannoukos  
Dawn Ciulla  
Jim Bochicchio

## Funding

NHGRI, Broad, Illumina

## DFCI

Kim Stegmaier

## Health Univ. de Barcelona

Jaume Mora

## 10X Genomics ❖

The entire team

❖ yes we are hiring  
fun in California

